Analysis and Design of an Algorithm for Optimization of Frequent Pattern Mining

SHORT SYNOPSIS

Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

Manav Rachna International Institute of Research and Studies
(Deemed to be University)

by

Anshu Singla
Registration number: 18/PhD/62

Under the supervision of

Dr. Parul Gandhi
Associate Professor, Faculty of Computer Applications

FACULTY OF COMPUTER APPLICATIONS

MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH & STUDIES

SECTOR-43, SURAJKUND –DELHI ROAD, FARIDABAD – 121001

FACULTY OF COMPUTER APPLICATIONS

Certificate

I hereby declare that the short synopsis "Analysis and Design of an Algorithm for Optimization of Frequent Pattern Mining" being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Applications under faculty of Computer Applications of Manav Rachna International Institute of Research & Studies, Faridabad during the academic year 2019-20. It is a bonafide record of original work carried out under the supervision of Dr. Parul Gandhi and has not been submitted in full or part to any other university or institute for degree or diploma. Further, it is certified that short synopsis is plagiarism free and proof has been submitted.

29/01/2020

Anshu Singla
Registration Number:
18/Ph.D/62

29/01/2020

Dr. Parul Gandhi
Associate Professor
Faculty of Computer Applications

(Chairman, DRC) 29/01/2020

Chairperson, CDP 29/01/2020

FACULTY OF COMPUTER APPLICATIONS

Certificate

I hereby declare that the short synopsis "Analysis and Design of an Algorithm for Optimization of Frequent Pattern Mining" being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Applications under faculty of Computer Applications of Manav Rachna International Institute of Research & Studies, Faridabad during the academic year 2019-20. It is a bonafide record of original work carried out under the supervision of Dr. Parul Gandhi and has not been submitted in full or part to any other university or institute for degree or diploma. Further, it is certified that short synopsis is plagiarism free and proof has been submitted.

Anshu Singla
Registration Number:
18/Ph.D/62

Dr. Parul Gandhi
Associate Professor
Faculty of Computer Applications

(Chairman, DRC)

Chairperson, CDP

ABSTRACT

Data Mining is used for extraction of useful patterns from huge datasets that may help in decision making. Frequent Pattern Mining (FPM) is very focused research area in data mining over a decade. FPM is used in association rule mining and other data mining tasks. Frequent Patterns are patterns those occur very frequently in dataset. As there are existing algorithms for FPM like Apriori, Fp Growth, Eclat, P-eclat, SDFEM, and FP-ICA. In Apriori algorithm , there is need of multiple scanning of datasets for generation of candidate itemsets . But in Fp-Growth algorithm dataset is scanned two times to form Fp-tree. In Fp Growth huge memory consumption is done to hold the fp tree for Frequent Pattern Mining. In this era of big data, these limitations are more prominent to deal with large datasets. A comparative study of some existing FPM algorithms and their variations has been analyzed to find the limitations of these algorithms. Based on the study a new algorithm will be proposed which should be more scalable and efficient in terms of time and memory. This algorithm will be developed by pruning irrelevant itemsets to deal with Frequent Pattern Mining. In addition to this, proposed algorithm will be framed into parallel and distributed environment to balance the workload of single node by using hadoop technology. Efficiency and scalability both parameters will be improved by using distributed and parallelization on bigdatasets. Real time datasets will be used to validate proposed algorithm and compare the result of proposed algorithm with the existing Frequent Pattern Mining algorithms.

Keywords: Frequent Pattern Mining (FPM), Fp-growth, Fp-tree, Association, Pruning, hadoop, Parallelization, Big datasets.

CONTENTS

S. No.	Description	Page No.
1	Introduction	5-7
2	Literature Review	7-9
3	Description of Broad Area	9
	3.1 Frequent Patterns	10
	3.2 Association analysis	10
	3.3 Interesting and useful Patterns	10-11
	3.4 Efficiency and Scalability	11-12
4	Research Gap	12
5	Objectives of the Study	12
6	Methodology to be adopted	12-13
7	Expected Outcome of the Research	13
	References	13-16

1. INTRODUCTION

Data Mining is a technique used for extraction of meaningful and useful information from collection of huge datasets. Useful and required information can't get directly from the database as it seems. Data mining is a complex process which is the collection of some simple steps. The various steps of data mining process are data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation .These steps should be followed in the same order. With the quick increment in measurement and open databases in different zones, it seems very simple to consider how to mine the data from enormous measure of data. Knowledge extraction in database targets to find the significant data from enormous collection of data. With the mining of knowledge in data bases, large data bases fill in as a reliable source for knowledge and the mined data can be helpful to numerous applications, for example, query processing, decision making, etc. Different clients consider data mining as another utilized term known as KDD (Knowledge Discovery and Data Mining). Figure 1 shows the different steps of knowledge extraction in data mining.

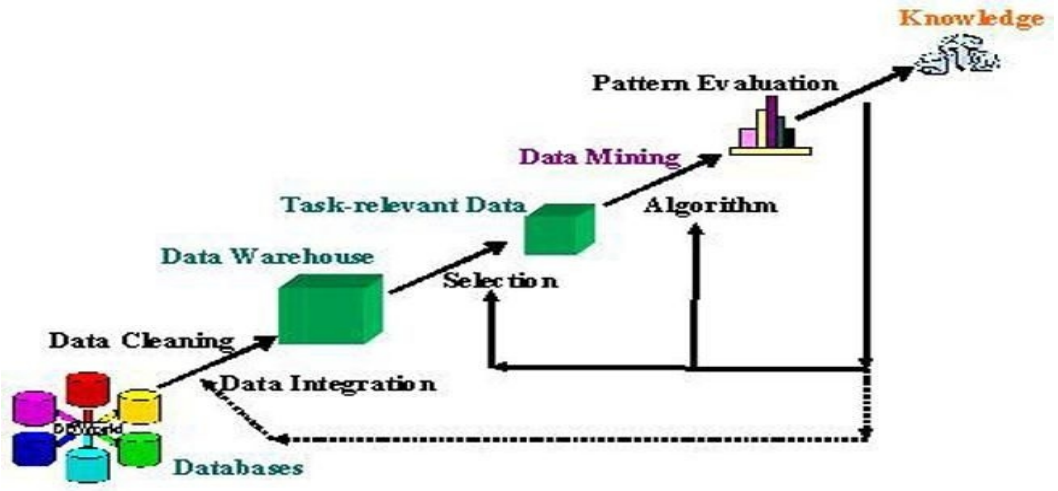


Figure 1 : Knowledge Discovery and Data Mining (Source: Han and Kamber 2001).

Data Mining Process

Data Mining processes are divided into seven steps:

a) Data Cleaning : Data gets cleaned in the data preparation. Real time data is always incomplete ,inconsistent and noisy. Many times data available does not include attributes we required. For example, if we need of demographic data of clients and the attribute of gender or address is not available, then there is no need of this type of data, means data is incomplete .Many times data contain some errors and outliers. Like age attribute with value 260 is an error. The data present in database may be inconsistent. For example, the name of an employee is Ramesh might be stored as Ramesh Kumar in different tables or documents. In that case the data is inconsistent. If the data provided for data mining is not clean, noiseless and consistent, the data mining results will not be accurate and useful. Data cleaning involves a number of techniques including fill the missing values by manual process etc. As a result, data received from data preprocessing steps will be cleaned data.

b) Data Integration: Data is combined and integrated from many resources to get a single database in data integration step. Data stored in various forms in various locations. Data can be stored in various forms like database, word files, sheets, forms, documents, data warehouses, websites and more. This is very tricky and complex task since data is available at different locations in various formats. Metadata is very helpful in data integration. Another problem in this step is data redundancy. Data Integration provides reliable data to maximum capacity by removing data redundancy.

c) Data Selection: As huge data of historical databases used for analysis. In data repository, data is available in large volumes as required. From this huge data, data

of interest is selected for specific purpose. Data selection is the process where data of interest is selected from huge volume of data sources.

d) Data Transformation: Data is being transformed in other form of data. Data transformation can be done by using data normalization, data aggregation and data generalization. For eg, a data set is “5, -37, 110, 87, -48" can be changed to as "0.05, -0.37, 1.10, 0.87, -0.48" by using transformation. After data transformation, data mining can be proceed easily.

e) Data Mining: Data mining is very important and main process combination of various tidy processes which are applied to find Patterns from large data for knowledge representation. Data mining process contains a many tasks like association, classification, prediction, clustering, time series analysis and more.

f) Pattern Evaluation: Useful Interesting patterns are finding out in this step. These patterns are depending on some interesting measures. If any pattern is useful, easily understandable or valid, only then these patterns can be used.

g) Knowledge Representation: Knowledge or information extracted from data has the need to be presented in some applicable manner. There are many visualization techniques available for knowledge representation.

2. LITERATURE REVIEW

- D. Yan, W. Qu, G. Guo and X. Wang (2020) proposed PrefixFPM , A general programming framework that can be used to mine frequent itemsets. Prefix projection is used for utilizing CPU completely while divide and conquer approach is used for dividing workload.
- Titarenko,S.,Titarenko,V.,Aivaliotis,G. & Palczewski,J.(2019), the author developed an algorithm to solve time series problems category, of time stamp uncertainties and without time stamp uncertainties. The algorithm also solved time series with temporal constraints and without temporal constraints. Most widely used algorithm is SPAM for sequential pattern Mining and Robust SPAM are referred and used in this paper. The proposed algorithm is much better as it works faster 20 times by SPAM and 6000 times faster as compared to RobustSPAM.

- Whang(2019) proposed a new data analytic model for MISFP (multi item support frequent pattern) in hadoop based parallel algorithms to attain high efficiency by multi item support frequent pattern. Experimental results reduces the time complexity of parallel algorithm by 38% by using MISFP algorithm in distributed environment.
- MRQAR is first non exhaustive algorithm on MapReduce proposed by Martín in 2018.
- Another non-exhaustive algorithm is G3P-LSC a genetic programming language in MapReduce framework for Frequent pattern mining is proposed by Padillo in 2018.
- Caroro, A.,Sison,A.& Medina,R.P (2018) , the author proposed dual pruned Frequent Pattern growth in this paper. In first pruning, it removes itemsets having minimum support less than minimum support count. In second phase, it removes the nodes which are not satisfying minimum support count. He used modified anti-monotone for second step traversing and developed 2P-FP algorithm for pruning .
- According to Ajinkya Kalaskar, Vishali Barkade (2018), a technique proposed known as Fp-growth algorithm in which information is compressed to form Fp-tree and Fp-tree is used for frequent itemsets. Limitations of more memory and more run time in WSC Algorithm are overcome by Fp Growth Mechanism.
- Dawen Xiahow (2018) , the author discovered hadoop technology that provides better result as compared to MapReduced technique for spatiotemporal Frequent Patterns .In this paper a parallel frequent pattern growth named as MR-PFP algorithm is proposed. Hadoop technology is used instead of MapReduce for parallelization. The results show that MR-PFP is better than earlier Parallel FP-growth (PFP) algorithm to improve efficiency and scalability.
- Mohamed A. Gawwad, Mona F. Ahmed & Magda B. Fayek,(2017), the Author developed an algorithm by using the Buddy Prima algorithm . The Greatest Common Divisor (GCD) is also used in the algorithm. A new technique has been developed for parallelization of Fp growth algorithm. Candidate itemsets are not used for itemsets generation . MapReduce and Spark is used for big datasets to explore parallelization abilities in hardware instead on single node.
- Fan Zhang,Youan Xiao (2017) developed DGFP (Dynamic Grouping Frequent Pattern) aims load Balancing which is the problem in FP Growth .It groups the data set by using dynamic grouping. Load balancing method is proposed in this paper. Algorithm

proposed improves load balancing effectively and reduces the overall efficiency of cluster by 5 to 15%.

- Pei,B.,Wang,X.&Wang.F.(2016), in this paper author proposed an algorithm PNFP to find association rules by using existential probability of itemsets in transactions .PNFP is more efficient as it is parallelized algorithm by using MapReduce framework to deal with big datasets. It minimizes data replication and communication cost of frequent pattern mining.
- Masome Hoseini & Shahraki (2016), the author introduced an algorithm based on transactional database , which are updated every time ,there is need of incremental Association rule Mining algorithm .CanTree (Canonical Tree) is used instead of FP-tree in frequent pattern growth Algorithm which improves efficiency at high mininmum support.
- Min Chen, XueDong Gao(2015) , a new parallel Fp Growth algorithm is proposed by the author to run on cluster system. Projection method is used instead construction of fp tree to find Frequent itemsets. Problem of memory overflow is also solved in this paper. Any data mining task splitted into many subtasks and these subtasks are executed in parallel .After execution the aggregated result is the final result. By this way efficiency and scalability improves much better.
- Gupta and Garg (2011) extends the work on COFI algorithms and show that performance of COFI algorithms lowers in case of sparse dataset.
- El-Hajj and Zaiane (2003) proposed an algorithm based on pruning method. In pruning method data storage can be reduced by cutting down irrelevant itemsets from FP-Tree used in FP-Growth algorithms. These pruning algorithms are called Co-Occurrence Frequent Itemset (COFI) .COFI -tree construction works in top down mining technique. Memory consumption is less as compared to FP-Tree and hence less computation time is needed. Two types of implementations are there in COFI- Tree 1).Non- recursive method is used to generate all frequent patterns . 2).pruning method are used to remove non frequent itemsets, only frequent itemsets are remaining ,however threshold value of minimum support value consider low.

3. DESCRIPTION OF BROAD AREA

Data Mining is the extraction of useful patterns from huge databases. These patterns are used in market basket analysis, which bring revolutionary changes in creating revenue. Frequent patterns are used for association rule mining. Apriori algorithm is used earlier for association rule mining .Association rule mining is used to create interesting patterns.

3.1 Frequent Patterns

Frequent Patterns are the patterns that happen frequently in data. There are numerous types of frequent patterns. A lot of frequent itemsets occurs frequently in transactional databases which provide useful information for decision making. For example, pencil and eraser, which are frequently purchased together in markets by various buyers. A frequently occurring subsequence, for example, another frequent pattern that buyer, purchased a PC, along with computerized camera, and then a memory card, is a (Frequent) consecutive pattern. There are many different consecutive sequences of these frequently itemsets.

3.2 Association Rules

Association rules are used to find association between unrelated itemsets in transactional databases. These association rules are used for analysis and prediction of customer or buyer's interest .The association rules are used for market basket analysis which can give beneficial results and make revenue to the authorities. Support and Confidence are widely used for generating association rules. Support threshold and Confidence threshold are used to calculate frequent itemsets. Values of these thresholds are decided by user. For example: In 60% of the cases, when customers buy milk, they also buy sugar. This shows the association between milk and sugar. We can represent it as: milk->sugar| 60% association

3.3 Interesting and Useful Patterns

A Data Mining framework can possibly create thousands or even a huge number of examples, or rules.

Let A={a_1 , a_2, a_3,……,a_{d}} is collection of different item's in a market.

T= {t_1,t_2,t_3,t_4……t_n} are some transactions of the store. Any transaction is made up by collection of some items chosen from itemset A. In association analysis, k-itemset is collection of k items from itemset A. For example, any transaction t_i = {milk, bread, sugar} has 3 items then it is 3-itemset.Null itemset having zero items. Association analysis is used to find relationship between different items in voluminous data. Support and Confidence are two coins used to find interesting and useful patterns. Support determines how often any association rule is applicable to dataset. Confidence represents how frequently item appears in transaction. Strong association rules are selectively those rules having support and threshold value greater then user specific minimum support and minimum threshold values.

3.4 Efficiency and Scalability

Efficiency and scalability are constantly viewed as when contrasting information Mining calculations. As information sums keep on duplicating, these two elements are particularly basic. Efficiency and scalability of information Mining calculations: Data mining calculations must be productive and versatile so as to viably separate data from tremendous measures of information in numerous information archives or in powerful information streams. As it is, the running time of an information mining calculation must be unsurprising, short, and worthy by applications. Efficiency, scalability, execution, advancement, and the capacity to execute continuously are key criteria that drive the improvement of numerous new information mining calculations, parallel, appropriated, and gradual mining calculations. The huge size of numerous informational indexes, the wide storage of information, and the computational power of certain information Mining strategies are factors that affect the improvement of parallel and circulated information escalated mining calculations. Such calculations first segment the information into "pieces." Each piece is handled, in parallel, via looking for

designs. The parallel procedures may collaborate with each other. Distributed computing and group figuring, which use PCs in a cooperative manner to handle enormous scale computational data, are likewise dynamic research subjects in parallel information mining. The significant expense of certain information Mining forms and the gradual idea of information advance steady information mining, which consolidates new information refreshes without mining the whole information "without any preparation." Such techniques perform information alteration steadily to correct that is recently found.

4. Research Gaps

The literature survey discussed in section 2 provides the various algorithms for Frequent Pattern Mining. There are still gaps which are as follows.

I. Effective pruning method can be used by removing irrelevant patterns to improve efficiency and speed of Frequent Pattern Mining Algorithms.

II. Dual pruning based on constrained frequent pattern can be used to remove unnecessary load on memory to retain irrelevant frequent patterns , further enhance the efficiency of FPM algorithms in terms of space and memory.

III. An enhanced algorithm can be developed to overcome various issues like memory, efficiency and scalability in FPM algorithms.

IV. Since the databases are updated every time, so incremental Frequent Pattern Mining is required.

5. OBJECTIVES OF THE STUDY

In view of the limitations of the existing algorithm, the objectives of our study shall be

1. Analytical study of existing Frequent Pattern Mining Algorithms proposed by various researchers.
2. Identification of an effective Pattern pruning method to retain high quality Pattern.
3. Propose an algorithm for Frequent Pattern Mining framed into parallel and distributed environment to balance the workload of single node.
4. Validate the proposed algorithm using real datasets.

6. METHODOLOGY TO BE ADOPTED

In the first phase, we will identify factors that impacts the Frequent Pattern Mining and in the second phase pruning technique will be used for removing irrelevant itemsets effectively. Hadoop technology will be used to develop a new algorithm for Frequent Pattern Mining framed into parallel distributed environment to balance the workload of single node, .In Hadoop environment we will use real time dataset to validate our proposed algorithm and comparison with existing algorithms.

7. EXPECTED OUTCOME OF THE RESEARCH

Pruning of irrelevant itemsets will help in developing efficient algorithm for Frequent Pattern Mining. After pruning of irrelevant itemsets, resultant algorithm will be useful for FPM in big datasets. Proposed algorithm will consume less memory, time and cost as compared to existing Frequent Pattern Mining algorithms. Parallel and distributed file processing will help in less usage of CPU to read and write data form cluster.

REFERENCES

D. Yan, W. Qu, G. Guo and X. Wang(2020) "PrefixFPM: A Parallel Framework for General-Purpose Frequent Pattern Mining," 2020 IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, 2020, pp. 1938-1941.

Luna, José María .(2019) “Frequent Itemset Mining: A 25 Years Review.” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, volume 9, Issue June, 2019.

Chee, Chin-Hoong(2018) “Algorithms for Frequent Itemset Mining: a Literature Review.” Artificial Intelligence Review, Volume 52, Issue April, 2018, pp. 2603–2621.

Caroro,A.,Sison,A.& Medina,R.P (2018)“An Enhanced Frequent Pattern-Growth Algorithm with Dual Pruning using Modified Anti-Monotone Support” IEEE 2018. pp719-728 Volume 12, Issue January,2018.

Titarenko,S.,Titarenko,V.,Aivaliotis,G. & Palczewski,J.(2019) “Fast implementation of Pattern Mining algorithms with time stamp uncertainties and temporal constraints” Springer Open, Journal of Big Data . pp 1-24 ,Volume 2,Issue March,2019.

Kalaskar,A. & Barkade,V.(2018) “FP-Growth Policy Mining For Access Control Policies” IEEE pp49-57 Volume 11,Issue August 2018

Chee,C.H.,Jaafar,J.,Aziz,I.A.,Hasan,M.H.,Yeoh,W.(2018):Algorithms for Frequent itemset mining: a Literature review “Springer pp248-254 Volume 10,Issue April 2018.

Xia,D.,Lu,X.,Li,H.,Li,Y.&Zhang,Z.(2018)“A MapReduce-Based Parallel Frequent Pattern Growth Algorithm for Spatiotemporal Association Analysis of Mobile Trajectory Big Data” Hindawi Volume 2018. pp 211-218, Volume 5,issue March 2018.

Zhaopeng,P.,Peiyu,L.&Jing,Y.(2018) “An Improved FP-tree Algorithm for Mining Maximal Frequent Patterns” IEEE ,pp 69-78 Volume 2,Issue May 2018

Gawwad, M.A., Ahmed,M.F.,Fayek.M.B.(2017) “Frequent Itemset Mining for Big Data Using Greatest Common Divisor Technique”. Data Science Journal, pp. 1–10 Volume 12,Issue April 2017.

Shah,S.,Amjad,M.,Habeeb,A.,Faruqui.M & Shafi.M (2017) “Algorithms for Frequent Pattern Mining of Big Data” International Journal of Applied Engineering Research pp74-78 ,Volume 6,Issue June 2017.

Yuzhou,Z.,Jianyong,W.,Lizhu,Z.(2017) “Research and Improvement of Parallelization of FP Growth based on Spark” IEEE . pp567-578 ,Volume 12,Issue August 2017.

Yuzho,Z.(2017) “Parallel Frequent Pattern Discovery: Challenges and Methodology” TSINGHUA SCIENCE AND TECHNOLOGY ,pp 589-595, Volume 4,Issue July 2017.

Wenzhe,L.Qian,W. & Changzhen,H.(2017) “Mining Frequent Patterns for Item Oriented and Customer Oriented Analysis” IEEE , pp23-31, Volume 2, Issue February 2017.

Xiao,M.,Yin,Y.,Zhou,Y.,Pan,S.(2017) “Research on Improvement of Apriori Algorithm Based on Marked Transaction Compression” IEEE pp 56-60, Volume 5, Issue July 2017.

Pei,B.,Wang,X.&Wang.F.(2016)”Parallelization of FP-growth Algorithm for Mining Probabilistic Numerical Data based on MapReduce “ IEEE 2016 9th International Symposium on Computational Intelligence and Design. Pp 589-593,Volume 12, Issue November, 2017

Gatuha,G.,Jiang,T.(2016) “Novel Frequent Pattern Mining Algorithm based on Parallelization scheme”. International Journal of Engineering Research in Africa , pp 131-140 ,Vol. 23,Issue December 2016.

Hoseini,M.S.,Shahraki,M.N. &Neisiyani,B.S.(2015) “A New Algorithm for Mining Frequent Patterns in Can Tree” IEEE 2nd International Conference on Knowledge Based Engineering and Innovation Iran. Pp235-242, Volume 25,Issue august 2015.

Vu,L.,Alaghband,G.(2015) “A Load Balancing Parallel Method for Frequent Pattern Mining on Multi-core Cluster” Society for Modeling & Simulation International (SCS) pp568-575 ,Volume 22, Issue April 2015.

Jamsheela,O.,G,R.(2015) “Frequent Itemset Mining Algorithms :A Literature Survey” IEEE pp19-28 ,Volume 15,Issue July 2015.

Anastasiu,D.C.(2015)“Big Data Frequent Pattern Mining”. University of Minnesota, Twin Cities, U.S pp123-130, Volume 22, Issue June 2015.

Sandip,B.S.,Apurva,D.(2015) “Latest Survey On Frequent Pattern Mining: Mine The Frequent Patterns From Transaction Database” pp 214-222,Vol.4.No.1,Issue January 2015.

Deng,Z.,Lv,S.L(2014) “Fast Mining Frequent itemsets using Node sets” Elsevier pp585-592 ,Volume 12,Issue March 2014.

Hui-ling,P.,Yun-Xing,S.(2012) “A New FP-tree-based Algorithm MMFI for Mining the Maximal Frequent Itemsets” IEEE pp365-372, Volume 23, Issue May 2012

Tan,J.,BU,Y.,Yang,Bo.(2009) “An Efficient Frequent Pattern Mining Algorithm” 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery pp245-252, Volume 12, Issue September, 2009.

MANAV RACHNA
Vidyanatarikshah

Dr. O.P. Bhalla Central Library
Manav Rachna International Institute of Research and Studies
(A deemed-to-be-university under section 3 of UGC Act 1956)

SIMILARITY VERIFICATION CERTIFICATE

Title: Analysis and Design of an Algorithm for Optimization of Frequent Pattern Mining

Researcher: Anshu Singla Registration Number: 18/PhD/162

Supervisor: Prof. (Dr.) Parul Gandhi Department: FCA

This is to certify that the above mentioned Short Synopsis/~~Thesis~~ was scanned for similarity index through Turnitin Software on ~~29.01.2020~~ with submission ID 1241810597 and the similarity index was found to be 10%. The complete report running into 17 Pages, 2831 Word count with File Size 1798 KB is being forwarded to the Research Supervisor/DRC/DRB for review.

Date: 29.01.2020

Dr. Vikram Kr. Sharma
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International
Institute of Research & Studies
Faridabad

University Librarian

The complete report of the above Short Synopsis/~~Thesis~~ has been reviewed and the similarity index is below the accepted norms as per the University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations 2018. The Short Synopsis/~~Thesis~~ may be considered for further processing

Singla
Research Scholar

29/1/2020
Research Supervisor

Short Synopsis Anshu Singla Jan.29,2020

by Anshu Singla

Submission date: 29-Jan-2020 12:43PM (UTC+0530)
Submission ID: 1248110597
File name: short_synopsis_Anshu_Singla_jan.29,2020.docx (75.89K)
Word count: 2831
Character count: 16791

Analysis and Design of an Algorithm for Optimization of Frequent Pattern Mining

SHORT SYNOPSIS
by

Anshu Singla
18/PhD/62

Under the supervision of

Dr. Parul Gandhi

ABSTRACT

Data Mining is used for extraction of useful Patterns from huge Datasets that may help in decision making. Frequent Pattern Mining (FPM) is very focused research area in Data Mining over a decade. FPM is used in association rule Mining and other Data Mining tasks. Frequent Patterns are Patterns those occur very frequently in Dataset. As there are Existing Algorithms for FPM like Apriori, Fp Growth, Eclat, P-eclat, SDFEM, and FP-ICA. In Apriori algorithm ,there is need of multiple scanning of datasets for generation of candidate itemsets . But in Fp-Growth Algorithm Dataset is scanned two times to form fp-tree. In Fp Growth huge memory consumption is done to hold the fp tree for Frequent Pattern Mining. In this era of Bigdata, these limitations are more prominent to deal with large datasets. A Comparative study of some existing FPM Algorithms and their variations has been analyzed to find the limitations of these algorithms. Based on this study a new Algorithm will be proposed which should be more scalable and efficient in time and memory. This algorithm will be developed by pruning irrelevant itemsets to deal with Frequent Pattern Mining. In

Short Synopsis Anshu Singla Jan.29,2020

ORIGINALITY REPORT

10%	4%	5%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

7/01/2020

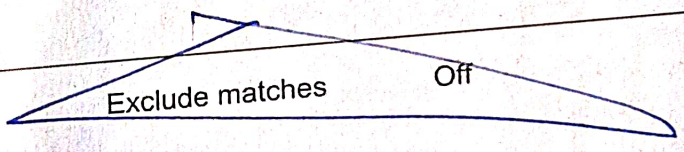
Dr. Vikram Singh Sharma
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International
Institute of Research & Studies
Faridabad

1	Submitted to University of Malaya Student Paper	2%
2	downloads.hindawi.com Internet Source	1%
3	data-flair.training Internet Source	1%
4	datascience.codata.org Internet Source	1%
5	www.tutorialspoint.com Internet Source	<1%
6	Submitted to iGroup Student Paper	<1%
7	www.scientific.net Internet Source	<1%
8	ijipc.com.ng Internet Source	<1%
9	Submitted to Middlesex University Student Paper	<1%

10	Submitted to University of Sunderland Student Paper	<1%
11	Linlin Jia, Laisheng Xiang, Xiyu Liu. "An Improved Eclat Algorithm Based on Tissue-Like P System with Active Membranes", Processes, 2019 Publication	<1%
12	Lecture Notes in Computer Science, 2016. Publication	<1%
13	bioinfopublication.org Internet Source	<1%
14	pure.hud.ac.uk Internet Source	<1%
15	Submitted to National Institute of Technical Teachers' Training and Research, Bhopal Student Paper	<1%
16	Submitted to Coventry University Student Paper	<1%
17	Submitted to University of Birmingham Student Paper	<1%
18	Khalid Iqbal, Xu-Cheng Yin, Hong-Wei Hao, Qazi Mudassar Ilyas. "Learning Bayesian Network leveled-structure from support based XML frequent itemsets", The 2013 International Joint Conference on Neural Networks (IJCNN),	<1%

2013
Publication

19	Submitted to Napier University Student Paper	<1%
20	Submitted to Jawaharlal Nehru Technological University Student Paper	<1%
21	Frequent Pattern Mining, 2014. Publication	<1%
22	Hao Jiang, Xu He. "An Improved Algorithm for Frequent Itemsets Mining", 2017 Fifth International Conference on Advanced Cloud and Big Data (CBD), 2017 Publication	<1%



Exclude quotes	On
Exclude bibliography	On